Methodology on Creating the U.S. Linked Retail Health Clinic (LiRHC) Database

by

Alice Zawacki
U.S. Census Bureau

Joey Marshall
U.S. Census Bureau

Donald Cherry
National Center for Health Statistics

Xianghua Yin
National Center for Health Statistics

Brian W. Ward
National Center for Health Statistics

CES 23-10 **March 2023**

Abstract

Retail health clinics (RHCs) are a relatively new type of health care setting and understanding the role they play as a source of ambulatory care in the United States is important. To better understand these settings, a joint project by the Census Bureau and National Center for Health Statistics used data science techniques to link together data on RHCs from Convenient Care Association, County Business Patterns Business Register, and National Plan and Provider Enumeration System to create the Linked RHC (LiRHC, pronounced "lyric") database of locations throughout the United States during the years 2018 to 2020. The matching methodology used to perform this linkage is described, as well as the benchmarking, match statistics, and manual review and quality checks used to assess the resulting matched data. The large majority (81%) of matches received quality scores at or above 75/100, and most matches were linked in the first two (of eight) matching passes, indicating high confidence in the final linked dataset. The LiRHC database contained 2,000 RHCs and found that 97% of these clinics were in metropolitan statistical areas and 950 were in the South region of the United States. Through this collaborative effort, the Census Bureau and National Center for Health Statistics strive to understand how RHCs can potentially impact population health as well as the access and provision of health care services across the nation.

Keyword*:* Retail health clinic locations, data linkage, probabilistic matching, County Business Patterns Business Register, health care

JEL Classification: C19, I1

---

* The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. Data Management System (DMS) number: 7529814. Disclosure Review Board (DRB) approval numbers: CBDRB-FY23- CES019-007. The corresponding author is Alice.M.Zawacki@Census.gov

I. Introduction

This methodological report details the construction of a nationally representative database of retail health clinic (RHC) locations operating throughout the United States from 2018 through 2020. This database, the Linked RHC (LiRHC, pronounced "lyric") database, was created by linking multiple datasets both internal and external to the U.S. Census Bureau and could be used to help improve understanding of the characteristics of RHCs and their surrounding geographic settings. Furthermore, LiRHC may allow analysts the ability to calculate estimates on the presence of RHCs and the characteristics of retailer(s) with an RHC. From this, estimates can be generated that could characterize the population and potentially other health care providers located in proximity to the RHC.

LiRHC's construction and a plan for future analytics is particularly timely because the landscape and types of health care settings where routine, preventive ambulatory medical care is provided has significantly changed over the last decade. Furthermore, the changing manners in which this care is delivered have been exacerbated in recent years by the ongoing COVID-19 pandemic. Historically, ambulatory care was typically provided in medically designated settings such as physician offices, federally qualified health centers, and at times hospital emergency departments. However, alternative settings have been increasingly utilized for needed and preventive ambulatory, or direct outpatient, care (RAND Corporation, 2016).

In its collection of RHC data from the retail and healthcare subsectors in the 2017 Economic Census, the U.S. Census Bureau defines a RHC as an in-store clinic with health care professional(s) who provides medical care (Basker, et al. 2019). An example of a RHC commonly given is a CVS Minute Clinic found inside a CVS retail location. This care may include various types of ambulatory medical care, including health screenings, immunizations, treatment of minor illnesses and injuries, and medication management. Between 2008 and 2015, there was a 36% decrease in visit rate for low-acuity conditions to hospital emergency departments, yet a 214% increase in the visit rate to RHCs (Poon, Schuur, and Mehrotra, 2018). It has been suggested that the popularity of RHCs has been influenced by factors such as convenient locations and hours, walk-in accessibility, short waiting times, and transparent pricing. It has subsequently been stated that such factors may make RHCs "a low-cost, high-quality, and convenient alternative to traditional primary care offices and emergency rooms" (Kaissi 2016). In contrast, others have expressed concern with the quality of RHC services, as well as the lack of a physician-patient relationship (RAND Corporation, 2016).

Research on RHCs is limited but growing. Investigation on the number of RHCs from 2010 to 2016 shows an increase of 66% from 1,224 to 2,036 and a number of quarterly entries and exits (Geddes and Schnell, 2022). In regards to patient utilization, several studies showing the increase in RHC and urgent care center (UCC) utilization in 2019 have been published by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS) (Black and Zablotsky, 2020; Black and Adjaye-Gbewonyo, 2021). Some findings included that among adult users: (a) women were more likely than men to have had at least one RHC/UCC visit in the past 12 months, (b) the use of RHCs/UCCs decreased as age increased, (c) non-Hispanic white adults were more likely than other race and Hispanic-origin groups to have had one or more RHC/UCC visits, and (d) RHC/UCC use increased as education level increased. While useful, this research is limited as it is unable to distinguish the use of UCCs from

RHCs. This distinction is important, as both UCCs and RHCs are distinct settings with different levels of use where 36.5% of U.S. adults (95.9 million) used UCCs for in-person care in 2022 compared with 30.3% of adults (79.6 million) who used RHCs (Leventhal 2022). Furthermore, it is also based on household interviews where adults self-report their use of RHCs/UCCs and does not include any provider or facility-level information. Thus, this lack of facility-level information does not allow the demographic, contextual, and business factors related to these RHCs to be examined.

In order to fill this gap, this methodology report describes procedures for LiRHC's creation that was derived by linking the U.S. Census Bureau's business data on retail trade, healthcare, and other industries with external data sources. The result was the creation of a RHC database that can be used to examine not only the characteristics of RHCs and retailers with RHCs but can be geographically matched with additional data sources to describe the populations living near RHCs and other available healthcare services located nearby.

In Section II, the sources used in the matching process to create LiRHC are described. One data source included is the Convenient Care Association Membership, which has data on both RHCs and other types of walk-in health care centers. The County Business Patterns Business Register is the second data source used, and contains a variety of information related to payroll, employment size, and business characteristics. Finally, the National Plan and Provider Enumeration System non-public use files were used in developing LiRHC, and these contain data on individual health care providers and non-individual organizations.

Section III details the matching methodology for linking these data sources for a RHC database. Specifically, four stages used are detailed.

- Stage 1 includes the preprocessing of the records.
- Stage 2 describes the matching of the Convenient Care Association data to the County Business Patterns Business Register data.
- Stage 3 details the additional linkage and matching of the National Plan and Provider Enumeration System to these aforementioned data sources.
- Stage 4 describes the postprocessing of the matched records.

Once the matching was completed, a quality assessment of the matching was performed that is described in Section IV. This includes the performance of benchmarking, use of match statistics, and the performance of manual review and quality checks. Section V uses the data resulting from the matching to provide a description of the geographic location of the RHCs contained in the database. Finally, the sixth and final section provides a brief conclusion and discusses some steps that are planned where LiRHC could be used to further describe and understand the areas and population characteristics surrounding the RHCs identified.

II. Data Sources

Convenient Care Association Membership

The Census Bureau's Center for Economic Studies (CES) purchased data[1] on membership in the Convenient Care Association (CCA), which can include both in-store RHCs as well as other types of convenient walk-in health care centers (e.g., UCCs).[2] These data were originally purchased to assist with validating and supplementing the Census Bureau's 2017 Economic Census data, which collected data on RHCs from the retail trade and health care sectors (Zawacki, 2020). In the current research project, these data are instrumental for creating LiRHC given their coverage of convenient care clinics.

Although the current LiRHC database is constructed with RHC locations in 2018-2020, CES purchased CCA datafiles for the years 2010-2020. These files include pulls of data for one or more months within each year, with the majority of RHC locations contained in all pulls within a year.[3] The variables in the files are generally consistent across all pulls and years; however, some differences are noteworthy. For example, to help illustrate differences in the STORE_NAME field, we use a well-known example, CVS MinuteClinic. In the 2010-2013 CCA datafiles, the STORE_NAME variable indicates the name of the *retailer*, which in our example would be CVS. In contrast, STORE_NAME in the 2014-2020 CCA datafiles includes the name of the RHC, which in this example is MinuteClinic. The annual CCA files include a phone number, as well as several address identifiers, including: CITY, STATE, ZIP_CODE, LATITUDE, and LONGITUDE. The 2016-2020 CCA files also contain COUNTY, COUNTRY, COUNTRY_CODE, and GEO_ACCURACY.[4,5] The original data files do not contain identifiers that disambiguate one unique RHC location from another; instead these are constructed after deduplication thereby enhancing these data for research. Described in Section III are the methods for matching the CCA records to the Census Bureau's business data, where these linkages rely primarily on the CCA name and address fields.

County Business Patterns Business Register

The Business Register (BR) is the data source that serves as the universe of all U.S. business establishments[6] and serves as the sampling frame for many Census Bureau surveys. As described by DeSalvo, Limehouse, and Klimek (2016), the BR is developed from several different information sources. The Internal Revenue Service provides line items from payroll and business income tax records. The Social Security Association (SSA) provides BR details on new businesses and organizational taxpayers,

[1] These data were purchased from AggData, which markets their expertise in locational data (www.aggdata.com).
[2] For more details on the Convenient Care Association, see www.ccaclinics.org.
[3] A total of 34 files were purchased with 1-5 pulls per year. The majority of RHCs appeared in each pull, but no single pull contained all RHCs.
[4] GEO_ACCURACY is created by the data providers to reflect the validity of the geographic detail such as rooftop verification or the location measure(s) are based the premise's rooftop or geometric center,
[5] The 2012 CCA file also contains store hours, while the 2016 file uniquely includes a second address field (ADDRESS_LINE_2).
[6] An establishment is the physical location where business activity takes place. A firm may own or operate one establishment (i.e., single unit establishment/firm) or more than one establishment (i.e., multi-unit establishment/firm).

and the Bureau of Labor Statistics (BLS) contributes state-level establishment data based on unemployment insurance programs to the BR. The Census Bureau's collection of business data in its Company Organization Survey and Economic Censuses also contribute to the BR's development and maintenance.

Section III below details the methods for identifying RHC locations and for using the edited BR files for publishing the County Business Patterns (CBP) reports. These CBP publications include estimates on establishment counts, payroll, and employment by county and industry. We use the CBP version of the BR files since they contain edited and published data that assists in the CCA-CBPBR matching procedures.

Several CBPBR measures are used in the matching procedures. For example, positive employment and payroll values as well as indicators for records used when publishing the Census Bureau's official tabulations assist with deduplicating CBPBR matches to a CCA record. The process also uses the North American Industry Classification System (NAICS) codes to assist with prioritizing matches to retailers generally affiliated with in-store RHCs, including pharmacies and drug stores, department stores, supermarkets, other non-convenience grocery stores, and warehouse clubs and supercenters. NAICS codes for health care providers offer information on those who may be involved in a RHC's operation.[7] Section III details the use of these key CBPBR measures during the linking processes.

In future planned analyses, the establishment's physical address from the annual CBPBR files will support characterizations of the populations and other health care providers located near RHC locations. The CBPBR location measures will also potentially assist with matches to external data sources. In addition, the available establishment, parent firm, and tax unit identifiers can support future planned linkages to additional Census Bureau business data to describe RHC organizations and characteristics.

National Plan and Provider Enumeration System (NPPES)

Under a Census Bureau agreement with the Centers for Medicare & Medicaid (CMS), this RHC project has access to the non-public National Plan and Provider Enumeration System (NPPES) files that contain Employer Identification Numbers (EINs). The NPPES files, also known as the National Provider Identifier (NPI) Registry, contain data on individual providers (Type 1 entities) and non-individual organizations (Type II entities). The NPI is a standard unique identifier for both health care providers and health plans and is used for administrative and financial transactions such as billing.

The primary reason for linking the NPPES' organizational entities to the matched CCA-CBPBR records is to help evaluate the quality of these matches by using additional NPPES measures. RHCs are often co-located with retailers such as pharmacies that bill Medicare and Medicaid for prescriptions, vaccinations, durable medical equipment, and/or other medical care-related items. For this reason, we expect many

[7] In future planned work using the 2017 Economic Census, respondents' self-designated 9-digit NAICS (621493-002) will also be reviewed for utility when distinguishing outpatient care facilities such as emergency or UCCs from RHCs.

linked CCA-CBPBR retail locations to match the NPPES organizations, which can help assist in distinguishing the RHCs from other types of clinics (e.g., UCCs).

CCA members can include different types of convenient walk-in clinics and centers aside from RHCs, such as UCCs. While the store or clinic name in the CCA records provides information suggesting whether or not the clinic is located within a retailer, and CCA-CBPBR matches help to identify retail locations, the CCA files do not contain measures distinguishing different types of clinics. For each NPPES health care organization, up to fifteen entries for each of the following is provided: (a) provider taxonomy codes, (b) taxonomy groups, and (c) primary taxonomy switch codes. These codes can help to distinguish RHCs from other types of clinics and to evaluate the quality of CCA-CBPBR matches by using the NPPES taxonomy for UCCs (i.e., 261QU0222X).

The NPPES files contain additional measures for health care organizations that help with evaluating the quality of the CCA-CBPBR matches. The non-public NPPES datafile used in identifying RHC locations includes the clinic's EIN, which references the tax unit.[8] Although the EIN does not necessarily correspond with a single retail establishment location with a RHC, it provides other information on firm ownership when assessing the quality of the CCA-CBPBR-NPPES matching that is described below. Other helpful NPPES measures include the name of the provider organization as well as the name and type code for an "other organization." Similar to the CBPBR measures for a business's physical address for the establishment's location, the NPPES files contain details on the organization's business practice location including a first- and second-line address, city, state, ZIP code, telephone, and fax number.

Additional NPPES measures with the provider's NPI enumeration date, along with any deactivation and reactivation dates, can also assist with possible closures of retailers who no longer bill or have a lapse in billing for health care services. This can be assessed relative to establishments (not) matching to the CBPBR. The NPPES files also contain measures indicating sole proprietors, whether the organization is a subpart to another business, the parent organization's legal business name, and flags for institutional providers and Medicare Part B suppliers.

III. Matching Methodology

Because the CCA records lack identifiers common to the CBPBR, they are linked using street addresses for the clinic's physical location. However, this presents a challenge as the street addresses from these different data sources often contain minor lexical variations. Consequently, the linkage process used in this research project utilized probabilistic matching built around the *similarity* of the address fields, which include street, city, and state. Address matches are especially fraught because the differences between valid and invalid matches are often miniscule; for instance, a next-door neighbor often has an identical address except for one differing digit of the address number.

Other fields common to both data sources, such as business names, were less useful because the 2018-2020 CCA records usually named the RHC whereas the CBPBR records generally listed the name of the

[8] EINs uniquely identify firms with one establishment (i.e., single-unit establishment/firm) and establishments within multi-unit firms that file taxes using a unique EIN for one of their establishments. However, some multi-establishment firms may file taxes using the same EIN for multiple locations.

retailer to which the clinic was associated. CCA records were first matched to the CBPBR and subsequently matched to the NPPES. The sections below describe the four stages in the matching processes.

Stage 1: Preprocessing Records

For all three constituent datasets – the CCA, the CBPBR, and the NPPES – we prepared the address fields for matching by applying basic text standardization such as setting all characters in lowercase and removing special characters. We also removed a set of common address identifiers such as "street" and "boulevard" (and their variations such as "st" and "blvd"). These terms can be regarded as address-related "stop words" that appear frequently and convey little lexical insight. Additionally, two records may contain different variations of a stop word (e.g., "st" versus "street"), which would present an unnecessary string mismatch between the two address records. In a hypothetical example such as "123 Main Street" versus "123 Main St," removing "St" from one record and "Street" from the other effectively standardizes the two addresses.

Through a process of iterative refinement, the most successful matching trials were those that used a concatenation of street address and city name. Occasionally, records from one of the input data sources would include the city name in the street address field. The intrusion of a city name into the address field sometimes changed the address string so drastically that the algorithm struggled to find a matching record. The simplest and most effective solution was to concatenate the address and city names for all records before attempting a match.

As noted above, in some instances invalid matches are often due to an issue where, for instance, a next-door neighbor often has an identical address except for one differing digit of the address number. To combat this "next-door neighbor problem" (in our case, a situation in which an establishment in the CCA would match to a next-door neighbor in the CBPBR), we enforced a perfect match on address number but allowed a probabilistic match on the rest of the address string. In other words, we tolerated minor lexical variations between data sources in the text of an address, but we required that the address *number* matched exactly between records.

The final preprocessing step involved deduplicating the CCA records, which was required since the CCA datafiles were created repeatedly during different months of the same year. We identified records as duplicates if their street address matched another record exactly.[9] Duplicates were removed; however, some information (such as the year and month for each CCA record) was retained and collapsed into the "primary" (i.e., non-duplicate) record such that no information was lost during this process.

[9] This deduplication step was deterministic – an exact address match identified a duplicate record. In some cases, duplicate records had very similar addresses with minor lexical variations that, to a human reviewer, were obvious duplicates. Those duplicate records are eventually identified via probabilistic deduplication and manual review, which are described in the section on stage 4.

Table 1 presents a fictitious example of how records from the CCA were preprocessed. This example demonstrates how duplicate address fields were initially seen in the datafile prior to processing. This table then shows how the record's address appeared after deduplication.

Stage 2: Linking the CCA to the CBPBR

The core of the matching procedure was built around the open-source *fastLink* R library, a parallelized implementation of the Fellegi-Sunter probabilistic record linkage model (Enamorado, Fifield, and Imai, 2019). To further optimize the matching procedure, the CCA and CBPBR data were split into state-year data extracts so that each state-year block (e.g., CCA State-A 2018; CBPBR State-A 2018) could be matched independently and simultaneously. Batches of state-year blocks were processed in parallel on one of the Census Bureau's research computing clusters.

The CCA-CBPBR matching process included eight passes, which are summarized in Table 2. In general, earlier passes such as pass 1 were computationally less resource-intensive than later passes (e.g., pass 5), which were tuned for less obvious address matches. These passes occurred sequentially, and cases were removed from the queue after a high-quality match was made. In this way, computing resources were reserved for the most difficult matches processed in the later passes.

The first matching pass linked CCA state-year blocks to CBPBR state-year records and was designed to match the those that were considered the least difficult or ambiguous. Some CCA records matched to multiple CBPBR records. As indicated in Table 2, these were deduplicated by *fastLink*, which selects and returns the most likely match out of a pool of possible match candidates. Because records are blocked on state and year, and because match deduplication is handled by *fastLink*, pass 1 is computationally less resource-intensive than later passes. In other words, pass 1 was designed to get the "easiest" matches out of the queue so that computing resources were not used unnecessarily in the proceeding passes. In pass 1, *fastLink*'s parameters were tuned to require a perfect match for address number. The concatenated address and city field (demonstrated in Table 1) was added to *fastLink*'s "stringdist.match" and "partial.match" arguments to allow probabilistic matching with the default thresholds.

The second pass was more resource-intensive than pass 1 and was designed to match records in mixed-use retail settings. A common problem we experienced was that RHCs were frequently located with other retail establishments (such as supermarkets, restaurants, hair salons, and banks) that all shared the same street address. [10] Because *fastLink* matches only on street address, all the collocated CBPBR establishments had the same probability of matching to a CCA record. To overcome this challenge, we tuned *fastLink* to return all of the likely match candidates in pass 2. Here *fastLink*'s arguments were tuned the same as pass 1, with the exception that *fastLink* was allowed to return duplicate matches.

The next step involved calculating the match quality scores using CBPBR criteria other than street address (Table 3). For instance, match candidates received a higher quality score if the name of the organization contained a keyword such as "clinic," if the establishment listed nonzero employment or payroll, or if the CBPBR's NAICS code indicated a probable industry for an RHC. Each match candidate

[10] A retailer with an in-store RHC may also be located at the intersection of two streets, and one street may be recorded in one data source and the cross street listed in the other source. We do not address this in these matches.

from the CBPBR began with a score of 0, and points were added for each of the listed criteria, for a possible maximum score of 23 points. Matches were retained and removed from subsequent passes if they had a quality score higher than 14, indicating high confidence that the CCA record matched to the correct CBPBR establishment.[11] This strategy effectively disentangled RHCs from other establishments at the same street address.

In sum, the key difference between pass 1 and pass 2 is that in pass 1 likely matches from the CBPBR were deduplicated and selected by *fastLink* itself. This was less resource-intensive than pass 2, where all possible matches from the CBPBR were returned, match quality scores were calculated, and the match candidate with the highest match quality score was chosen. Pass 1 was often unable to find a high-quality match for RHCs in mixed-use retail settings where multiple establishments were collocated at the same street address. However, pass 1 did quickly move the computationally "easiest" matches out of the queue, freeing up computing resources for the resource-intensive pass 2.

Later passes were all variations of passes 1 and 2. Passes 3 and 4 were identical to passes 1 and 2, but with one key difference: records from the CCA were allowed to search for a match in the previous year of the BR. For example, in passes 1 and 2 records from the 2020 CCA files were matched to establishments in the 2020 CBPBR. The CCA records that were unable to find a match in passes 1 and 2 cascaded to passes 3 and 4, where they were allowed to search the 2019 CBPBR for a match.

Passes 5-8 follow the same general structure as passes 1-4 but without the state-level blocking constraint. We observed that in a small minority of cases, discrepancies in the input data led to some records falling into the wrong state blocks based on the other available address, city, and/or zip code information. If cases were unable to find a high-quality match in passes 1-4, they cascaded to passes 5-8, in which records from the CCA were allowed to search all CBPBR states for a match. We took several measures to reduce computational burden in the absence of a state-level block, including: (a) only records unable to find a high-quality match in passes 1-4 were put through passes 5-8; (b) CCA cases were still subset by state (only one state-year from the CCA was matched at one time), although the CBPBR dataset was not restricted by state; and (c) CBPBR cases were subset down only to street addresses with an address number that also occurred in at least one CCA record. These measures greatly reduced the number of pairwise evaluations that would have occurred from trying to match the full universe of CCA and CBPBR records.

Stage 3: Linking the CCA-CBPBR to the NPPES

After the eight CCA-CBPBR matching passes were completed, the composite dataset was matched to the 2018 NPPES. Like the CCA and CBPBR, the NPPES data were also split by state, but not by year given the use of only the 2018 records. NPPES preprocessing was nearly identical to the steps described above in Stage 1, including concatenation of the city names to street addresses.

Unlike the CBPBR, NPPES records generally did not contain multiple establishments at the same street address. The NPPES also did not contain industries such as banks or restaurants, which could signal a

[11] A quality score of 14 was the decided cutoff for separating "good" from "poor" matches based on an examination of the results using an iterative process with manual reviews.

poor match to the CCA. This eliminated the need for the multiple passes used in the CCA-CBPBR matching described in stage 2. Rather, linking the CCA-CBPBR composite dataset to the NPPES required only two matching passes. In the first pass, records were blocked by state, and street addresses from the CCA and NPPES records were probabilistically linked. In the second pass, the state block was removed. In both passes, match deduplication and selection were determined by *fastLink.* By way of comparison, the two CCA-CBPBR-NPPES matching passes were nearly identical to passes 1 and 5 in the CCA-CBPBR matching methodology.

Stage 4: Postprocessing of the Matched Records

After matching each state-year block of the CCA records to the CBPBR and the NPPES, a series of postprocessing steps began with a final round of deduplication. This was necessary because the same RHC locations may operate in more than one year from 2018 to 2020. Additionally, the same RHC may be observed multiple times in the CCA file(s) with minor lexical variations in street address (e.g., "123 Main" versus "123 E Main"), which the deterministic deduplication step in preprocessing would have overlooked.

For these reasons, Stage 4 began by appending the 2018, 2019 and 2020 matched records together and performing a probabilistic or "fuzzy" deduplication step. The *fastLink* package and the Fellegi-Sunter model (Enamorado, Fifield, and Imai, 2019) was applied to deduplicate these by probabilistically matching the list of records to itself. Those that have a high probability of matching to more than one record are tagged as likely duplicates, which were then confirmed with a manual review. We dropped the duplicate records but retained some information from these, such as the year and month when each street address was observed. This information was collapsed into the "primary" (i.e., non-duplicate) record in a manner similar to the example described in Table 1.

After removing duplicates and unmatched CCA records, the resulting LiRHC database included 2,000 RHC locations operating at anytime during 2018 to 2020. Ninety-five percent of the RHCs in the CCA files matched to CBPBR establishments with a high-quality match per the following quality assessments.

IV. Quality Assessments

Various assessments of quality were performed on LiRHC with the 2018-2020 RHC locations after linking the CCA records to the CBPBR, and then subsequently to the NPPES.[12] First, the CCA records were reviewed following the deduplication process to eliminate RHC locations that might have been in multiple CCA files within a single year (See Stage 1 in Section III). These duplicates were eliminated through the iterative machine learning processes (described in Section III) as well as through manual checks and reviews.

[12] Quantitative findings from some of these quality assessments are presented when possible; however, some were deemed as being unable to be released given their disclosure risks.

Next, we benchmarked the number of deduplicated CCA records and the CCA-CBPBR matched RHC records to publicly available counts. Statista, which references its research services and surveys, provides free online access to 2019 RHC counts including those for the most common RHC operators.[13] Statista reports that there were 1,949 U.S. RHCs in 2019, and represents an estimated target for constructing the 2018-2020 database with RHC locations. An additional benchmark is available from Merchant Medicine data, which reports 2,008 RHCs existed as of January 1, 2017.[14] These online resources also provide RHC counts for the major operators, such as CVS, Walgreens, Kroger, Walmart, and Brooks Eckerd/Rite Aid. This allowed us to benchmark not only to the total RHC count, but also to the number of RHCs operated by these major retailer operators.

In addition to benchmarking, match statistics were also reviewed to help assess the quality of the linking processes. These included tabulations of the additive quality matching score assigned to each CCA-CBPBR record and the pass number for the resulting match. As shown in Table 4, the resulting quality scores for the matches ranged from 10-21 with 81% of the matches having a score of ≥18.[15] Matches with the lowest quality matching score were also manually reviewed for a final drop/keep decision. Employing the eight machine learning passes (described in Section III), the vast majority (92%) of the matched RHC locations were identified in the first two passes (Table 5). The high quality of the match rates was also confirmed by producing and reviewing crosstabulations of the assigned CCA retailer name[16] by the matched CBPBR business name.

After benchmarking and use of match statistics, manual review and quality checks were also performed on a random sample of 500 matched RHC locations. The team reviewed CCA locations matching to more than one CBPBR record, which sometimes resulted from different formatting of addresses as described above. Manual checking was also performed to review the NAICS and the NPPES taxonomies for the matched RHC locations. Some CBPBR locations matching to the CCA records had an unexpected NAICS code from industries unassociated with retail trade locations or healthcare services. Some matches were made to health care providers rather than retail locations suggesting the providers' involvement with the RHC's clinical services. After matching the CCA-CBPBR matches to the NPPES, the team reviewed the NPPES taxonomies to help distinguish UCCs from RHCs.

To allay concerns about selection bias for the unmatched vs. matched CCA records, a preliminary review was also conducted on the geographic distribution of the CCA records by their matched/unmatched status.

[13] See "Locations with the most retail clinics U.S. 2019 | Statista."
[14] This 2017 count is published online (Drug Channels: Retail Clinic Check Up: CVS Retrenches, Walgreens Outsources, Kroger Expands) and based on Merchant Medicine's database.
[15] As noted earlier, by design the maximum quality match score was 23. However, no final assigned scores exceeded 21.
[16] In the 2018-2020 CCA records, STORE_NAME reflects the RHC name and occasionally references the retail firm. When missing, the matching process assigned retailer names to these RHCs based on public records that supported crosswalks between RHC and retailer names. Some retailers may operate a RHC within another retailer's location. For example, CVS purchased and began operating RHCs in Target stores and later rebranded these as MinuteClinics. Firm name assignments take this possibility into account.

V. Retail Health Clinic Locations[17]

Following the matching processes and quality assessments, the final 2018-2020 LiRHC database has approximately 2,000 mutually exclusive observations. To describe their geographic distribution, these RHCs were geolocated in counties by using CBPBR or CCA latitude and longitude measures for each RHC site. A spatial join was performed between the point data and the 2018 vintage Census Cartographic Boundary File[18] to obtain the county Federal Information Processing Standards (FIPS) code for each RHC location and to then assign county-level metropolitan statistical area (MSA)/non-MSA classifications obtained from the 2013 NCHS urban-rural classification scheme for counties (Ingram and Franco, 2014).[19] As shown in Table 6, 97% of these locations are in MSAs. This is consistent with earlier work showing 88.4% of clinics were in an urban area (Rudavsky and Mehrotra, 2010).

Figure 1 shows the distribution of RHC locations by the four U.S. Census regions and nine divisions.[20,21] Of the total RHCs shown in this figure, 950 locations or almost half (47.0%) are located in the South region, with the majority (550 RHCs) found in the South Atlantic division. The Midwest region follows with 600 RHC locations, with the East North Central division containing twice the number of RHCs found in the West North Central division (i.e., 400 vs. 200 RHCs, respectively). These counts are consistent with other data sources showing the majority of RHCs are operated in southern and midwestern states.[22] As shown in Figure 1, the Northeast region has 240 RHC locations, with the majority in the Middle Atlantic division (150 RHCs). The fewest RHC locations are found in the West region (230 RHCs), with 150 RHCs located in the Mountain division and 80 in the Pacific division.

VI. Conclusion and Next Steps

Retail health clinics (RHCs) are a relatively new type of health care setting and developing the understanding of the role they play as a source of ambulatory care in the United States is important. In order to better understand RHCs across the United States, the Census Bureau and National Center for Health Statistics set to use data science techniques to link together three independent data sources to develop the LiRHC database, which provides a listing of RHCs in the United States. The sources used to create this database included the CCA, CBPBR, and NPPES. To successfully complete this merge, a multi-stage matching methodology was developed and implemented. This include first preprocessing the data in the CCA and CBPBR and was followed by linking the CCA and CBPBR using a series of eight sequential matching passes that assessed and performed various linkages. Next, the matched CCA-CBPBR data were then linked to the NPPES using a series of two passes. Finally, postprocessing procedures were completed to round out the matching methodology.

---

[17] These counts for RHC locations are rounded to meet disclosure avoidance rules.
[18] Available at https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html.
[19] For the counties in our analysis, the binary metro/non-metro classifications are identical between the 2013 NCHS classification scheme and the September 2018 delineation file from the Census Bureau.
[20] The figure shows a total of 2,020 RHC locations, which differs from the total of 2,000 RHCs, and is due to rounding.
[21] See Terms and Definitions (census.gov) for a listing of states in each census region and division.
[22] See Retail Health Clinic Locations in US - Location Analysis (scrapehero.com).

With this multi-stage matching methodology performed, we performed a quality assessment to ensure that the matches were of an adequate quality to be considered valid. This included utilizing multiple techniques, such as benchmarking, a statistical assessment, and finally manual review and quality checking. LiRHC tracked well with the benchmark sources, and the match statistics found that 81% of these matches were of high quality. Finally, the geographical distribution of the RHC locations was performed. Of the resulting linked 2,000 RHCs in LiRHC, 97% of these clinics were in a MSA and 950 were found in the South region of the United States. This distribution also was similar to the distributions seen elsewhere.

Through this collaborative effort focused on RHCs, the Census Bureau and NCHS strive to understand how RHCs can potentially impact population health as well as the access and provision of health care services across the nation. Thus, future projects with approved access to LiRHC can use these data to better understand the role of RHCs in health care provision in the United States. For example, using characteristics of the populations surrounding RHC locations, as well as the availability and distribution of other types of health care providers and settings near the RHCS in LiRHC would allow for a better understanding of how these RHCs may be utilized, and their potential to serve as a source of care for these nearby locations. Thus, using contextual data from the geographic areas in which RHCs are located could help enhance this understanding.

In addition, only the years 2018-2020 were used in this analysis. Performing the same matching and linkage using previous years of data (e.g., 2010-2017) could provide even more information on RHCs. For example, by assigning a unique identifier to RHCs, these data can examine entry and exit over time. Furthermore, the linkage of additional data sources, such as the Economic Censuses, the Services Annual Survey (SAS), and/or the Longitudinal Business Database (LBD) could increase the robustness of the LiRHC database and subsequently enhance its utility by adding business measures on employment size, payroll, and organizational structure. While the LiRHC database created in the current joint Census Bureau/NCHS project is a useful tool in understanding the role of RHCs in health care across the United States, continuing to build upon this foundation will only increase this knowledge further.

References

Basker, Emek, Randy A. Becker, Lucia Foster, T. Kirk White, and Alice Zawacki, 2019. "Addressing Data Gaps: Four New Lines of Inquiry in the 2017 Economic Census," Working Papers 19-28, Center for Economic Studies, U.S. Census Bureau.

Black Lindsey I., and Benjamin Zablotsky. 2020. "Urgent Care Center and Retail Health Clinic Utilization Among Children: United States, 2019." NCHS Data Brief, 393: 1-7.

Black Lindsey I., and Dzira Adjaye-Gbewonyo D. 2021. "Urgent Care Center and Retail Health Clinic Utilization Among Adults: United States, 2019." NCHS Data Brief, 409: 1-7.

DeSalvo, Bethany, Frank F. Limehouse, and Shawn D. Klimek. 2016. "Documenting the Business Register and Related Economic Business Data." U.S. Census Bureau, Center for Economic Studies Working Paper Series CES 16-17.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records." American Political Science Review, 113(2):353-371.

Fein, Adam J. 2019. The 2019 Economic Report on U.S. Pharmacies and Pharmacy Benefit Managers. Philadelphia, PA: Drug Channels Institute.

Geddes, Eilidh and Molly Schnell. 2022. "The Expansionary and Contractionary Supply-Side Effects of Health Insurance." October 2022.

Ingram Deborah D., and Shelia J. Franco. 2014. "2013 NCHS Urban–Rural Classification Scheme for Counties. Vital and Health Statistics, 2(166):1-73.

Kaissi, Amer. 2016. "Health Care Retail Clinics: Current Perspectives." Innovation and Entrepreneurship in Health, 31:47-55.

Leventhal, Rajiv. 2022. "Primary Care Practices Lose Patients to Alternative Sites of Care." Insider Intelligence's Digital Health Briefing, December 2. https://www.insiderintelligence.com/content/primary-care-practices-lose-patients-alternative-sites-of-care.

Poon, Sabrina J., Jeremiah D. Schuur, and Ateev Mehrotra. 2018. "Trends in Visits to Acute Care Venues for Treatment of Low-acuity Conditions in the United States From 2008 to 2015." JAMA Internal Medicine, 178(10):1342-1349.

RAND Corporation. 2016. "The Evolving Role of Retail Clinics." Research Brief, No. RB-9491. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_briefs/RB9491-2.html.

Rudavsky, Rena and Ateev Mehrotra. 2010. "Sociodemographic Characteristics of Communities Served by Retail Clinics." *Journal of the American Board of Family Medicine* 23 (1): 42-48.

Zawacki, Alice. 2020. “Evaluating New Content on Retail Health Clinics in the 2017 Economic Census.” Internal presentation on October 21, 2020 to the Business Research Area, Center for Economic Studies, U.S. Census Bureau. *Forthcoming* Census Bureau Technical Note.

Table 1: Fictitious Example of the Processing of CCA Records Performed in Stage 1 of the Matching Methodology

Duplicate CCA Records Before Processing

Address	City	Year|month_observed	Store_name
123 Main St.	Anywhere Town	201801	Clinic x
123 Main St.	Anywhere Town	201803	Clinic x
123 Main St.	Anywhere Town	201812	Clinic x

Deduplicated CCA Record

Perfect match field	Probabilistic match field		
Address_number	*Address_clean*	*Year|month_observed*	*Store_name*
123	123 e main anywhere town	201801, 201803, 201901	Clinic x

Table 2: Matching Passes for linking Convenient Care Association and County Business Patterns Business Register data

Pass	*Blocking*	*CBPBR match deduplication*	*CBPBR year*	*Goal*
1	Year, state	fastLink	Same as CCA	Gets the computationally "easiest" matches out of the queue
2	Year, state	By match quality score	Same as CCA	
3	Year, state	fastLink	CCA minus 1	Tries matching to the previous CBPBR year if necessary
4	Year, state	By match quality score	CCA minus 1	
5	Year	fastLink	Same as CCA	Passes 5-8 are identical to passes 1-4 but without state-level blocking
6	Year	By match quality score	Same as CCA	
7	Year	fastLink	CCA minus 1	
8	Year	By match quality score	CCA minus 1	

Source. Based on authors' merging of the 2018-2020 Convenient Care Association data (CCA) with the 2017-2020 County Business Patterns Business Register (CBPBR).

Table 3: Match Quality Scores for linking Convenient Care Association and County Business Patterns Business Register data

Criteria	*Score*
CBPBR name references a known RHC or contains a keyword such as "clinic" or "health"	+5
All numbers in address match (e.g., "123 Main, suite 500" = 123500)	+5
Address number match (e.g., "123 Main, suite 500" = 123)	+3
CBPBR name contains the name of a known grocer or major retailer	+3
CBPBR NAICS code* in expected industry	+3
CBPBR establishment reported positive payroll	+1
CBPBR establishment reported positive employment	+1
CBPBR establishment is an active business	+1
CBPBR establishment was tabulated in the LBD (positive data quality indicator)	+1

* Includes North American Industry Classification System (NAICS) codes starting with 621, 622, 623, 445110, 446110, 452210, 45231[23] or 452311.
RHC Retail health clinic
LBD Longitudinal Business Database

Source. Based on authors' merging of the 2018-2020 Convenient Care Association data (CCA) with the 2017-2020 County Business Patterns Business Register (CBPBR).

[23] This 5-digit code was included and may result in matches to other general merchandise stores (e.g., auto parts, home furnishings, etc.). This code will be removed in future iterations of the matching process. Current matches to these general merchandisers, where RHC locations are unexpected, were manually reviewed and addressed.

Table 4. Percentage Distribution of Retail Health Clinic Matches by Quality Score

Match Quality Score (0-100)	*Percent*
High (75-100)	81
Moderate to Low (0-74)	19

Source. Based on authors' merging of the 2018-2020 Convenient Care Association data, the 2017-2020 County Business Patterns Business Register, and the 2018 National Plan and Provider Enumeration System data from the Centers for Medicare and Medicaid Services.

Table 5. Matches by Pass Number

Pass Number	*Percent*
1 or 2	92
3-8	8

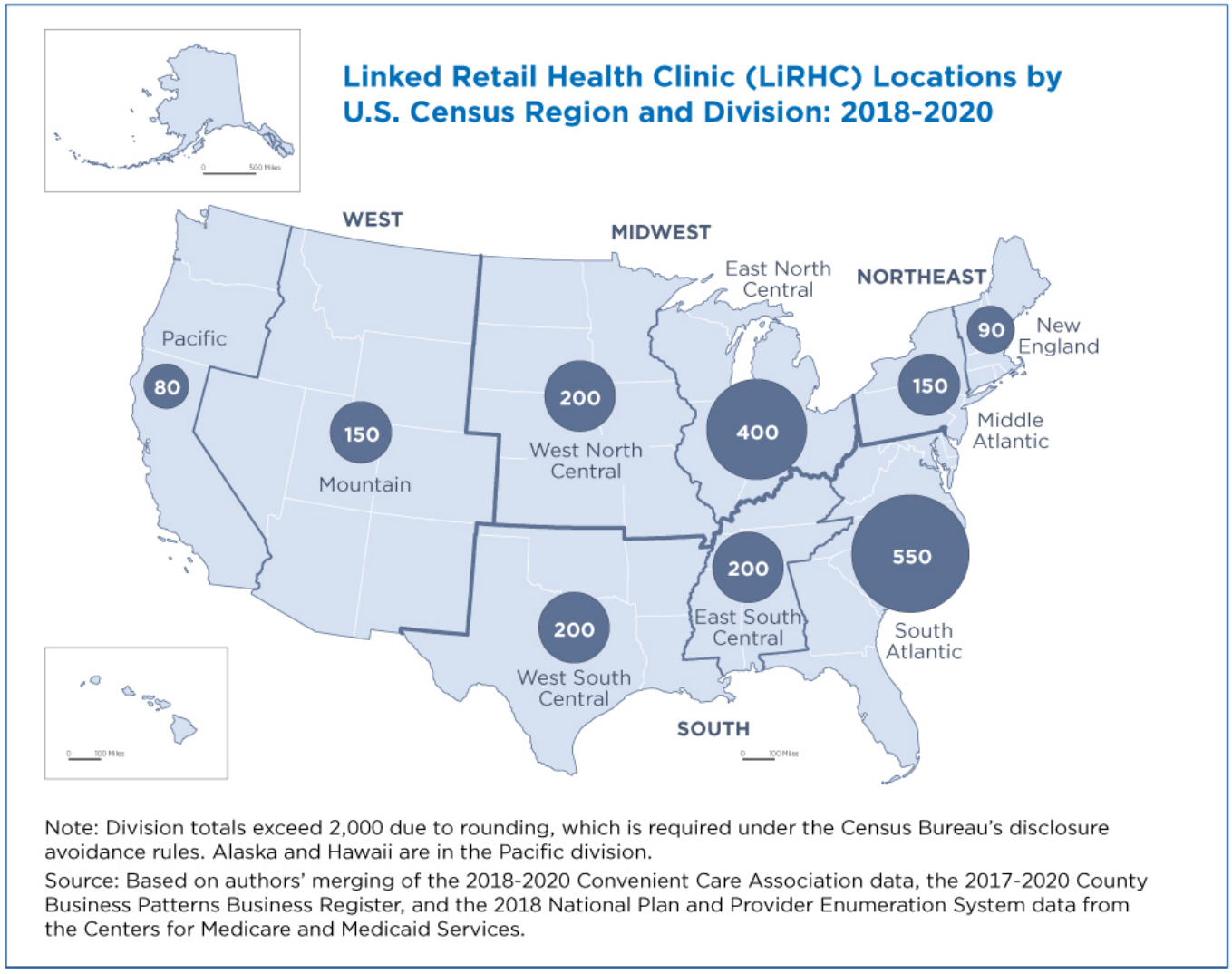
Source. Based on authors' merging of the 2018-2020 Convenient Care Association data, the 2017-2020 County Business Patterns Business Register, and the 2018 National Plan and Provider Enumeration System data from the Centers for Medicare and Medicaid Services.

Table 6. Linked retail health clinic (LiRHC) locations by metropolitan statistical area status, 2018-2020 (N=2,000)

Metropolitan Statistical Area	*Percent*
Yes	97
No	3

Source. Based on authors' merging of the 2018-2020 Convenient Care Association data, the 2017-2020 County Business Patterns Business Register, and the 2018 National Plan and Provider Enumeration System data from the Centers for Medicare and Medicaid Services.

Figure 1. Linked retail health clinic (LiRHC) locations by U.S. Census Region and Division, 2018-2020 [24]



Note: Division totals exceed 2,000 due to rounding, which is required under the Census Bureau's disclosure avoidance rules. Alaska and Hawaii are in the Pacific division.

Source: Based on authors' merging of the 2018-2020 Convenient Care Association data, the 2017-2020 County Business Patterns Business Register, and the 2018 National Plan and Provider Enumeration System data from the Centers for Medicare and Medicaid Services.

[24] The authors thank Kevin Hawley from the Census Bureau's Geography Division for creating this map.